清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

第七讲:多元线性回归分析

回归分析是数据分析中最基础也是最重要的分析工具，绝大多数的数据分析问题，都可以使用回归的思想来解决。回归分析的任务就是，通过研究自变量X和因变量Y的相关关系，尝试去解释Y的形成机制，进而达到通过X去预测Y的目的。

常见的回归分析有五类：线性回归、0-1回归、定序回归、计数回归和生存回归，其划分的依据是因变量Y的类型。本讲我们主要学习线性回归。

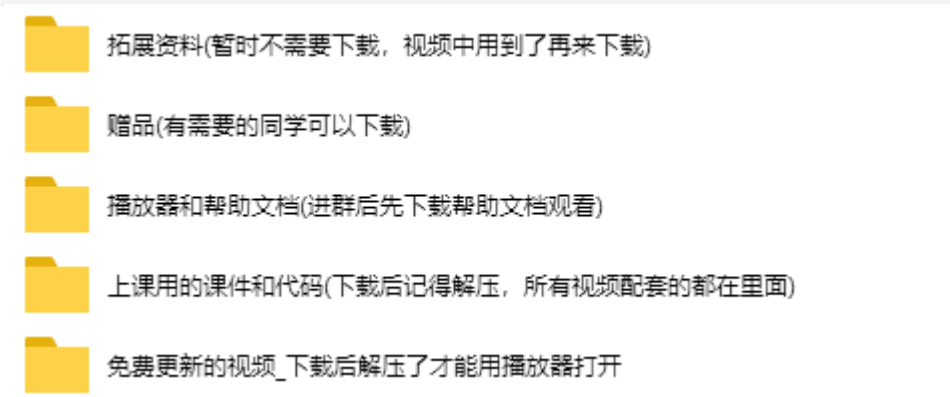
数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

温馨提示

（1）视频中提到的附件可在**售后群的群文件**中下载。

包括**讲义、代码、我视频中推荐的资料**等。



（2）关注我的**微信公众号《数学建模学习交流》**，后台发送**“软件”**两个字，可获得常见的建模软件下载方法；发送**“数据”**两个字，可获得建模数据的获取方法；发送**“画图”**两个字，可获得数学建模中常见的画图方法。另外，也可以看看公众号的历史文章，里面发布的都是对大家有帮助的技巧。

（3）**购买更多优质精选的数学建模资料**，可关注我的微信公众号《数学建模学习交流》，在后台发送**“买”**这个字即可进入店铺进行购买。

（4）视频价格不贵，但价值很高。单人购买观看只需要**58元**，和另外两名队友一起购买人均仅需**46元**，视频本身也是下载到本地观看的，所以请大家**不要侵犯知识产权**，对视频或者资料进行二次销售。

回归的思想

回归分析：研究X和Y之间相关性的分析。

三个关键词

1. **相关性**
2. **Y**
3. **X**

注：关于回归的很多观点，我引用了王汉生老师的《数据思维》，强烈推荐数据分析、统计等专业的同学阅读。

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

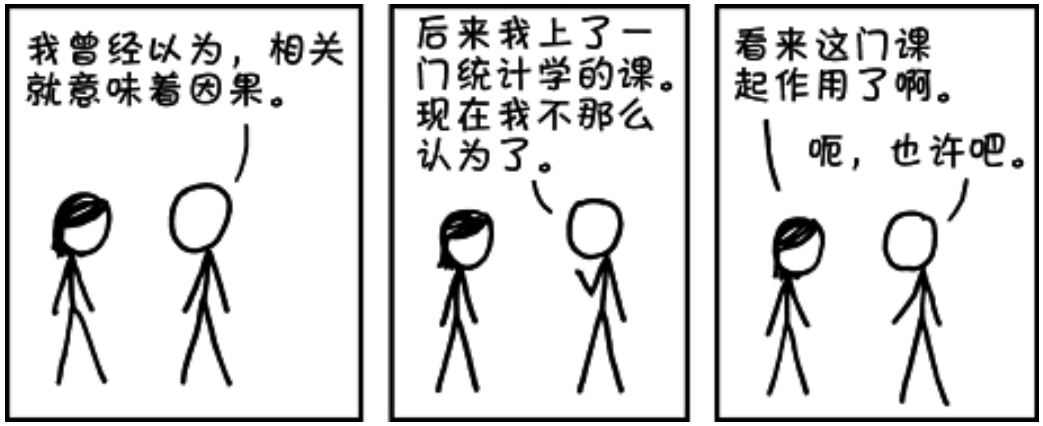
第一个关键词：相关性

统计数据表明：游泳死亡人数越高，雪糕卖得越多
（游泳死亡人数和雪糕售出量之间呈显著正相关）

可以下结论：吃雪糕就会增加游泳死亡风险吗？
（因为吃雪糕，所以游泳死亡风险增加了）

数学建模学习交流

相关性 ≠ 因果性



在绝大多数情况下，我们没有能力去探究严格的因果关系，所以只好退而求其次，改成**通过回归分析，研究相关关系**。

听起来比较悲观？其实不是的。为什么？因为，这个退而求其次的方案，比你瞎拍脑袋好多了去了。

数学建模学习交流

第二个关键词是：Y

Y是什么？俗称**因变量**。取义，因为别人的改变，而改变的变量。

在实际应用中，Y常常是我们需要研究的那个核心变量。

（1）经济学家研究经济增长的决定因素，那么Y可以选取GDP增长率**（连续数值型变量）**。

（2）P2P公司要研究借款人是否能按时还款，那么Y可以设计成一个二值变量，Y=0时代表可以还款，Y=1时代表不能还款**（0-1型变量）**。

（3）消费者调查得到的数据（1表示非常不喜欢，2表示有点不喜欢，3表示一般般，4表示有点喜欢，5表示非常喜欢）**（定序变量）**。

（4）管理学中RFM模型：F代表一定时间内，客户到访的次数，次数其实就是一个非负的整数。**（计数变量）**

（5）研究产品寿命、企业寿命甚至是人的寿命（这种数据往往不能精确的观测，例如现在要研究吸烟对于寿命的影响，如果选取的样本中老王60岁，现在还活的非常好，我们不可能等到他去世了再做研究，那怎么办呢？直接记他的寿命为60+，那这种数据就是截断的数据）**（生存变量）**

数学建模学习交流

第三个关键词是：X

Y是**因变量**(因为别人的改变，而改变的变量)。
而X是用来解释Y的相关变量，所以X被称为**自变量**。
当然，另一套定义方法是：X为解释变量，Y为被解释变量。

回归分析的任务就是，通过研究X和Y的相关关系，尝试去解释Y的形成机制，进而达到通过X去预测Y的目的。

例题：下表是1990-2007年中国棉花单产与要素投入的表格，请用回归的方法指出哪个要素投入是最重要的要素？

年份	单产 kg/公顷	种子费 元/公顷	化肥费 元/公顷	农药费 元/公顷	机械费 元/公顷	灌溉费 元/公顷
1990	1017.0	106.05	495.15	305.1	45.9	56.1
1991	1036.5	113.55	561.45	343.8	68.55	93.3
1992	792.0	104.55	584.85	414	73.2	104.55
中间1993-2004年的数据						
2005	1122	449.85	1703.25	555.15	402.3	358.8
2006	1276.5	537	1888.5	637.2	480.75	428.4
2007	1233	565.5	2009.85	715.65	562.05	456.9

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

0-1回归的例子

变量	变量说明
SUCCESS	借款是否成功，成功记为1
DEFAULT	获得借款后是否违约，违约记为1
LNAMOUNT	取对数后的借款金额
INTEREST	借款利率
MONTHS	借款的期限，共有6个选择： 3, 6, 9, 12, 18, 24月
INCOME	1表示月收入超过1万元，0表示不超过1万元
HOUSE	有房产则记为1，否则记为0
CAR	有车产则记为1，否则记为0
CREDIT	借款人的信用评级，1表示评级高，0表示评级低
WORKTIME	参加工作的时长，1表示工作时长在3年及以上
MARRED	婚姻状况：已婚记为1，未婚记为0
AGE	借款人的年龄
EDUCATION	本科及以上学历的借款人记为1，低于本科学历记为0

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

回归分析的使命

使命1：回归分析要去识别并判断：**哪些X变量是同Y真的相关，哪些不是。**统计学中有一个非常重要的领域，叫做“变量选择”。（逐步回归法）

使命2：去除了那些同Y不相关的X变量，那么剩下的，就都是重要的、有用的X变量了。**接下来回归分析要回答的问题是：这些有用的X变量同Y的相关关系是正的呢，还是负的？**

使命3：在确定了重要的X变量的前提下，我们还想**赋予不同X不同的权重，也就是不同的回归系数，进而我们可以知道不同变量之间的相对重要性。**

这就是回归分析要完成的三个使命：

第一、识别重要变量；

第二、判断相关性的方向；

第三、要估计权重（回归系数）。

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

回归分析的分类

类型	模型	Y的特点	例子
线性回归	OLS、GLS（最小二乘）	连续数值型变量	GDP、产量、收入
0-1回归	logistic回归	二值变量（0-1）	是否违约、是否得病
定序回归	probit定序回归	定序变量	等级评定（优良差）
计数回归	泊松回归（泊松分布）	计数变量	每分钟车流量
生存回归	Cox等比例风险回归	生存变量（截断数据）	企业、产品的寿命

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

数据的分类

横截面数据：在某一时点收集的不同对象的数据。

Cross Sectional Data

例如：

（1）我们自己发放问卷得到的数据
（2）全国各省份2018年GDP的数据
（3）大一新生今年体测的得到的数据

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

数据的分类

时间序列数据：对同一对象在不同时间连续观察所取得的数据。

Time Series Data

例如：

（1）从出生到现在，你的体重的数据（每年生日称一次)。

（2）中国历年来GDP的数据。

（3）在某地方每隔一小时测得的温度数据。

数学建模学习交流

数据的分类

面板数据：横截面数据与时间序列数据综合起来的一种数据资源。

Panel Data

例如：
2008-2018年，我国各省份GDP的数据。

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

不同数据类型的处理方法

数据类型	常见建模方法
横截面数据	多元线性回归
时间序列数据	移动平均、指数平滑、ARIMA、GARCH、VAR、协积
面板数据	固定效应和随机效应、静态面板和动态面板

建模比赛中，前两种数据类型最常考到；面板数据较为复杂，是经管类学生在中级计量经济学中才会学到的模型。

横截面数据往往可以使用回归来进行建模，我们通过回归可以得到自变量与因变量之间的相关关系以及自变量的重要程度。

时间序列数据往往需要进行我们进行预测，时间序列模型的选择也很多，大家需要选择合适的模型对数据进行建模。

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

数据的收集

因为提供数据的网站容易失效，所以大家可以直接在知乎上搜索“数据查找”来获取最新的数据网站。

上面的数据多半都是宏观数据，微观数据市面上很少
大家可以在人大经济论坛搜索
https://bbs.pinggu.org/

另外也可以自己学习爬虫
（1）Python等软件爬取（需要编程基础，实际学习起来不困难）
网易云课堂：零基础21天搞定Python分布爬虫
（2）傻瓜式软件爬取（八爪鱼）
免费分享给大家，请见本节配套课件的拓展资料文件夹

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

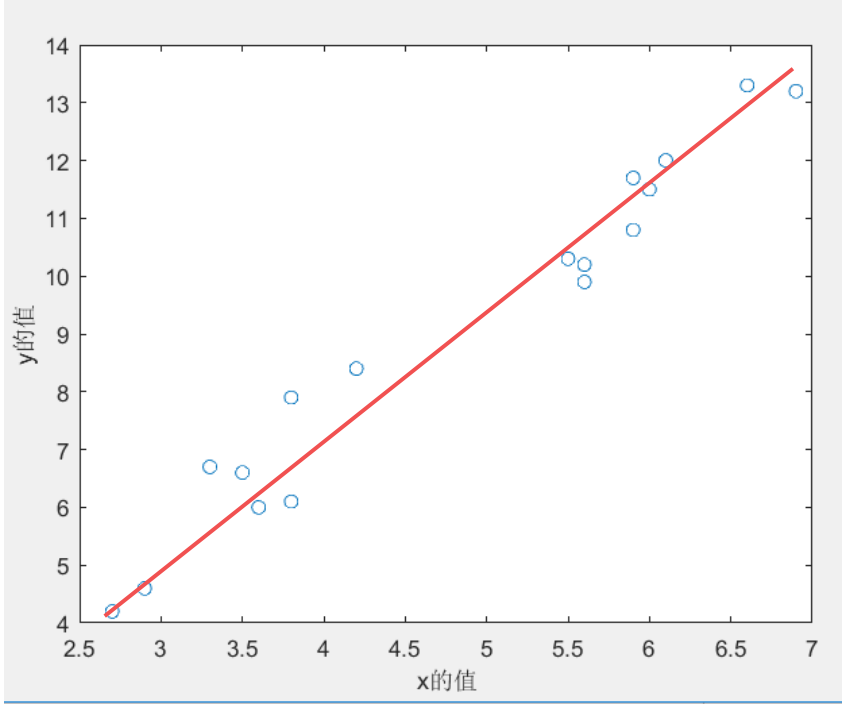
一元线性回归

一元线性函数拟合

设这些样本点为(x_i,y_i)，$i=1,2,\cdots,n$

我们设置的拟合曲线为$y=kx+b$

问题：k和b取何值时，样本点和拟合曲线最接近。



设这些样本点为(x_i,y_i)，$i=1,2,\cdots,n$，我们设置的拟合曲线为$y=kx+b$

令拟合值$\hat{y}_i = kx_i+b$

那么$\hat{k},\hat{b}=\underset{k,b}{arg\,min}(\sum_{i=1}^{n}(y_i-\hat{y}_i)^2)=\underset{k,b}{arg\,min}(\sum_{i=1}^{n}(y_i-kx_i-b)^2)$

令$L=\sum_{i=1}^{n}(y_i-kx_i-b)^2$，现要找$k,b$使得$L$最小。

（L在机器学习中被称为损失函数，在回归中也常被称为残差平方和）

一元线性回归模型

假设x是自变量，y是因变量，且满足如下线性关系：

$$y_i = \beta_0+\beta_1 x_i+\mu_i$$

β_0和β_1为回归系数，μ_i为无法观测的且满足一定条件的扰动项

令预测值$\widehat{y_i} = \widehat{\beta_0}+\widehat{\beta_1}x_i$

其中$\widehat{\beta_0},\widehat{\beta_1}=\underset{\beta_0,\beta_1}{arg\,min}(\sum_{i=1}^{n}(y_i-\widehat{y_i})^2)=\underset{\beta_0,\beta_1}{arg\,min}(\sum_{i=1}^{n}(y_i-\widehat{\beta_0}-\widehat{\beta_1}x_i)^2)$

$$\widehat{\beta_0},\widehat{\beta_1}=\underset{\beta_0,\beta_1}{arg\,min}(\sum_{i=1}^{n}(\widehat{\mu_i})^2)$$

我们称$\widehat{\mu_i} = y_i-\widehat{\beta_0}-\widehat{\beta_1}x_i$为残差

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

对于线性的理解

假设x是自变量，y是因变量，且满足如下线性关系：

$$y_i = \beta_0 + \beta_1 x_i + \mu_i$$

线性假定并不要求初始模型都呈上述的严格线性关系，
自变量与因变量可通过变量替换而转化成线性模型。

$$y_i = \beta_0 + \beta_1 \ln x_i + \mu_i$$

$$\ln y_i = \beta_0 + \beta_1 \ln x_i + \mu_i$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \mu_i$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \delta x_{1i} x_{2i} + \mu_i$$

使用线性回归模型进行建模前，需要对数据进行预处理。
用Excel、Matlab、Stata等软件都可以

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

回归系数的解释

$y_i = \beta_0 + \beta_1 x_i + \mu_i$ ， β_0和β_1为回归系数

假设x为某产品品质评分（1－10之间），y为该产品的销量，我们对x和y使用一元线性回归模型，

如果得到$\hat{y_i} = 3.4 + 2.3x_i$，如何解释我们估计出来的回归系数?

3.4：在评分为0时，该产品的平均销量为3.4

2.3：评分每增加一个单位，该产品的平均销量增加2.3

如果现在有两个自变量，x_1表示品质评分，x_2表示该产品的价格，那么我们可建立多元线性回归模型：

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \mu_i$，如果估计出来的回归等式为：$\hat{y_i} = 5.3 + 0.19x_{1i} - 1.74x_{2i}$

5.3：在评分为0且价格为0时，该产品的平均销量为5.3个（没现实意义）

0.19：在保持其他变量不变的情况下，评分每增加一个单位，该产品的平均销量增加0.19

-1.74：在保持其他变量不变的情况下，价格每增加一个单位，该产品的平均销量减少1.74

可以看到，引入了新的自变量价格后，对回归系数的影响非常大！！！

原因：遗漏变量导致的内生性

数学建模学习交流

内生性的探究

假设我们的模型为：$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$

μ为无法观测的且满足一定条件的扰动项

如果满足误差项μ和所有的自变量x均不相关，则称该回归模型具有外生性

（如果相关，则存在内生性，内生性会导致回归系数估计的不准确：不满足无偏和一致性）

回到刚刚那个例子：x为某产品品质评分（$1-10$之间），y为该产品的销量

我们建立的一元回归模型：$y_i = \beta_0 + \beta_1 x_i + \mu_i$，$\beta_0$和$\beta_1$为回归系数

那么在这个模型中：误差项μ_i包含什么？

包含了所有与y相关，但未添加到回归模型中的变量

如果这些变量和我们已经添加的自变量相关，则存在内生性

数学建模学习交流

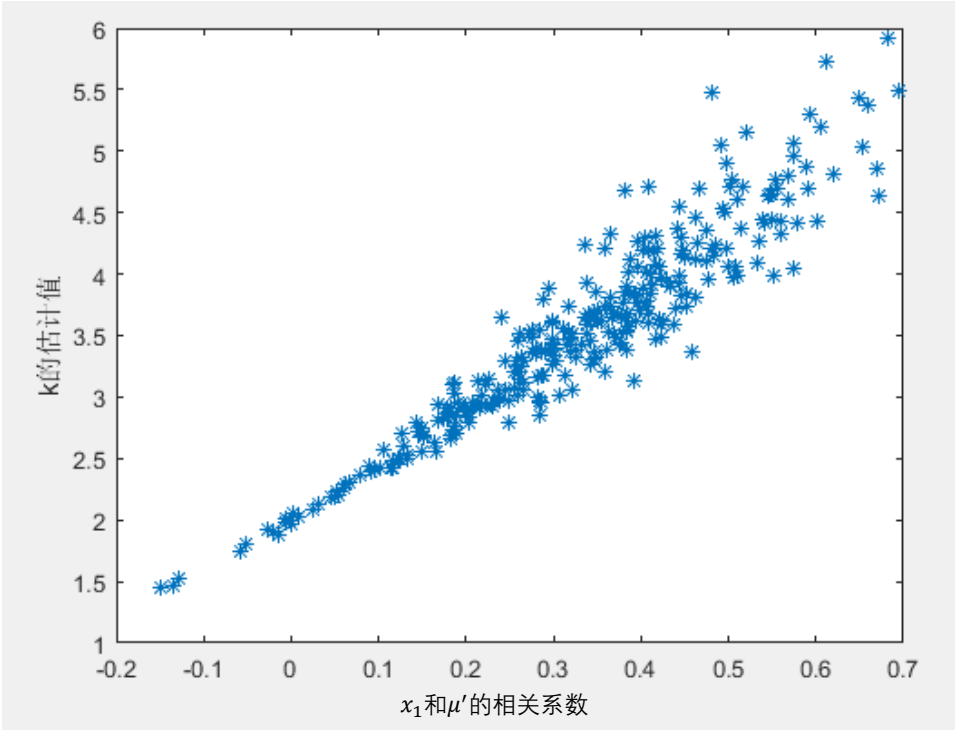
内生性的蒙特卡罗模拟

假设$y = 0.5 + 2x_1 + 5x_2 + \mu,\ \mu \sim N(0,1)$

如果x_1在$[-10,10]$上均匀分布

如果我们用一元线性回归模型：$y = kx_1 + b + \mu'$进行估计

试探究估计出来的k的大小与$\rho_{x_1,\mu'}$的关系



相关系数绝对值越大，代表内生性越大

数学建模学习交流

核心解释变量和控制变量

无内生性（no endogeneity）要求所有解释变量均与扰动项不相关。

这个假定通常太强，因为解释变量一般很多（比如，5-15个解释变量），且需要保证它们全部外生。

是否可能弱化此条件？答案是肯定的，如果你的解释变量可以区分为核心解释变量与控制变量两类。

核心解释变量：我们最感兴趣的变量，因此我们特别希望得到对其系数的一致估计（当样本容量无限增大时，收敛于待估计参数的真值）。

控制变量：我们可能对于这些变量本身并无太大兴趣；而之所以把它们也放入回归方程，主要是为了“控制住”那些对被解释变量有影响的遗漏因素。

在实际应用中，我们只要保证核心解释变量与μ不相关即可。

参考资料：再论OLS：核心变量与控制变量的区别

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

回归系数的解释

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \mu_i$$

$$\Downarrow$$

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{1i} + \widehat{\beta_2} x_{2i} + \cdots + \widehat{\beta_k} x_{ki}$$

$\widehat{\beta_0}$的数值意义一般我们不考虑，因为所有的自变量一般不会同时全为0。

$\widehat{\beta_m}\ (m=1,2,\cdots,k)$: 控制其他自变量不变的情况下，$x_{mi}$每增加一个单位，对$y_i$造成的变化。

实际上可以用数学中的偏导数来定义：$\widehat{\beta_m} = \dfrac{\partial y_i}{\partial x_{mi}}$

因此多元线性回归模型中的回归系数，也常被称为偏回归系数。

思考：回归模型$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} \ln x_i$中的$\widehat{\beta_1}$怎么解释？

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

什么时候取对数?

伍德里奇的《计量经济学导论，现代观点》里，第六章176-177页有详细的论述;

取对数意味着原被解释变量对解释变量的弹性，即百分比的变化而不是数值的变化;

目前，对于什么时候取对数还没有固定的规则，但是有一些经验法则:

（1）与市场价值相关的，例如，价格、销售额、工资等都可以取对数;

（2）以年度量的变量，如受教育年限、工作经历等通常不取对数;

（3）比例变量，如失业率、参与率等，两者均可;

（4）变量取值必须是非负数，如果包含0，则可以对y取对数ln(1+y);

取对数的好处: （1）减弱数据的异方差性（2）如果变量本身不符合正态分布，取了对数后可能渐近服从正态分布（3）模型形式的需要，让模型具有经济学意义。

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

四类模型回归系数的解释

1、**一元线性回归**：$y = a + bx + \mu$，x每增加1个单位，y平均变化b个单位；

例 A.1

线性住房支出函数

假定每月住房支出和每月收入的关系式是

$$housing = 164 + 0.27income \qquad (A.11)$$

那么，每增加 1 美元收入，就有 27 美分用于住房开支，如果家庭收入增加 200 美元，那么住房开支就增加 0.27×200=54 美元。图 A.1 描绘了这个函数的图形。

2、**双对数模型**：$lny = a + blnx + \mu$，x每增加1%，y平均变化b%；

例 A.5

常弹性需求函数

若 q 代表需求量，而 p 代表价格，并且二者的关系为

$$\log(q) = 4.7 - 1.25\log(p)$$

则需求的价格弹性是−1.25。粗略地说，价格每增加 1%，将导致需求量下降 1.25%。

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

四类模型回归系数的解释

3、**半对数模型**：$y = a + blnx + \mu$，x每增加1%，y平均变化b/100个单位;

例 A.7

劳动供给函数

假定一个工人的劳动供给可描述为

$$hours = 33 + 45.1 \log(wage)$$

其中，*wage* 为小时工资，而 *hours* 为每周工作小时数，于是，由式（A.30）得到，

$$\Delta hours \approx (45.1/100)(\%\Delta wage) = 0.451\%\Delta wage$$

换言之，工资每增加 1%，将使每周工作小时增加约 0.45 或略小于半个小时。若工资增加 10%，则 $\Delta hours = 0.451 \times 10 = 4.51$ 或约四个半小时，我们不宜对更大的工资百分数变化应用这个近似计算。

4、**半对数模型**：$lny = a + bx + \mu$，x每增加1个单位，y平均变化(100b)%。

例 A.6

对数工资方程

假设小时工资与受教育年数有如下关系：

$$\log(wage) = 2.78 + 0.094educ$$

应用式（A.28）就有

$$\%\Delta wage \approx 100\ (0.094)\Delta educ = 9.4\Delta educ$$

由此可知，多受一年教育将使小时工资增加约 9.4%。

以上例子来自：伍德里奇《计量经济学导论》第四版

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

特殊的自变量：虚拟变量X

如果自变量中有定性变量，例如性别、地域等，在回归中要怎么处理呢?

例如：我们要研究性别对于工资的影响(性别歧视）。

$$y_i = \beta_0 + \delta_0 Female_i + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \mu_i$$

$Female_i = 1$表示第i个样本为女性; $Female_i = 0$表示第i个样本为男性

核心解释变量：$Female$

控制变量：$x_m \ (m = 1, 2, \cdots, k)$

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

虚拟变量的解释

$$y_i = \beta_0 + \delta_0 Female_i + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \mu_i$$

$Female_i = 1$表示第i个样本为女性; $Female_i = 0$表示第i个样本为男性

$$E(y|Female = 1\text{以及其他自变量给定}) = \delta_0 \times 1 + C$$

$$E(y|Female = 0\text{以及其他自变量给定}) = \delta_0 \times 0 + C$$

$$E(y|Female = 1\text{以及其他自变量给定}) - E(y|Female = 0\text{以及其他自变量给定}) = \delta_0$$

所以δ_0可解释为:

在其他自变量给定的情况下，女性的平均工资与男性的平均工资的差异。

$$\widehat{wage} = -1.57 - 1.81 female + 0.572 educ + 0.025 exper + 0.141 tenure \qquad (7.4)$$
$$(0.72) \quad (0.26) \qquad (0.049) \quad (0.012) \quad (0.021)$$
$$n = 526, \ R^2 = 0.364$$

的 *educ*，*exper* 和 *tenure* 时，一个女性和一个男性在小时工资上的平均差距。如果我们找到受教育程度、工作经历和现职任期相同的一个女性和一个男性，那么平均来看，女性每小时比男性要少挣 1.81 美元。（要记住，这可是用 1976 年的工资水平来度量的!）

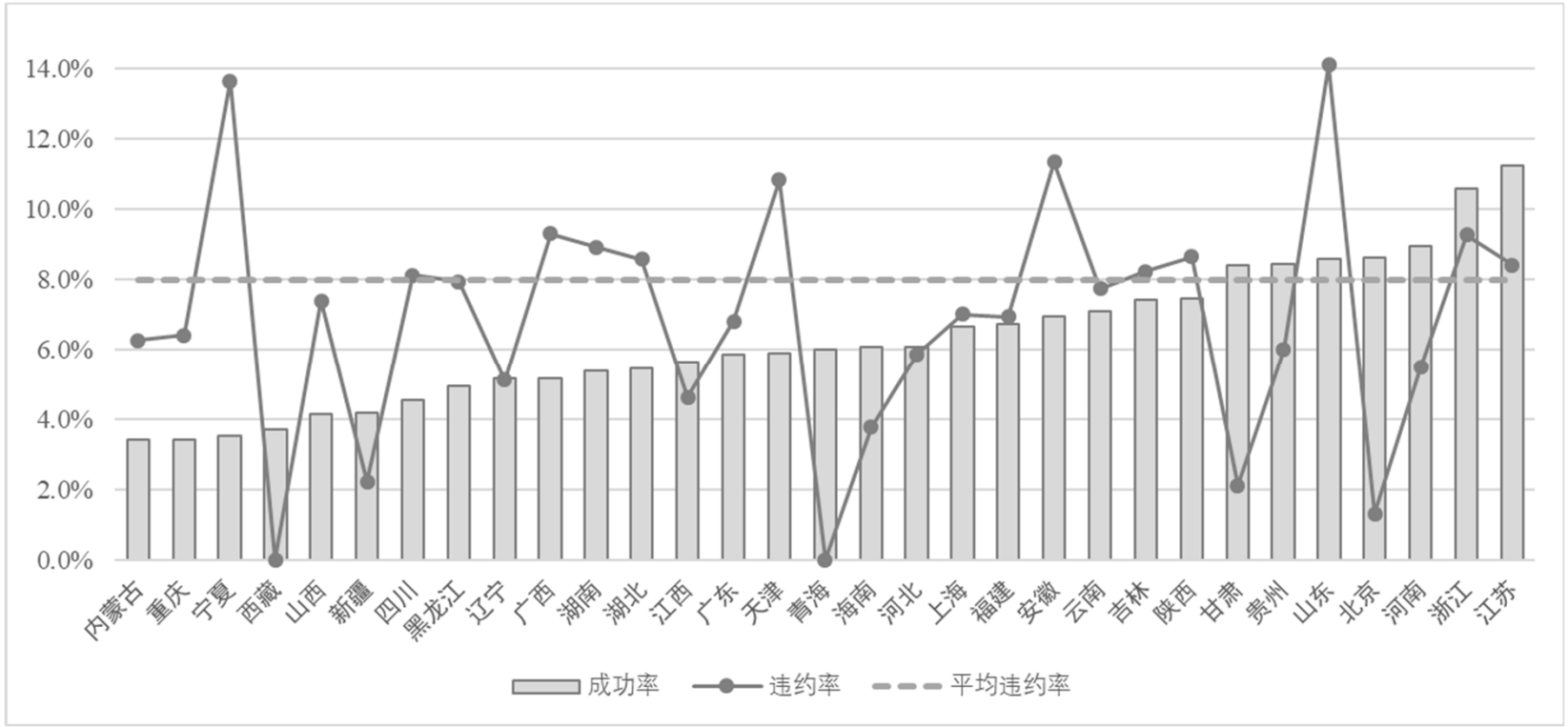
数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

多分类的虚拟变量设置

实证探讨我国P2P网络贷款中是否存在显著的地域歧视问题？



数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

多分类的虚拟变量设置

2.3 模型构建

为了检验 P2P 市场中是否存在地域歧视现象，本文构建了模型（1）：

$$SUCCESS_i = \alpha + \sum \beta_n \times Province_n + \lambda \times Controls_i + \varepsilon_i \quad (1)$$

就是扰动项μ_i
换了个符号而已

这里$SUCCESS_i$表示样本中第i个借款人是否获得借款，$Province_n$是省份的虚拟变量，剔除了港澳台三地后，还剩 31 个省份，本文设置内蒙古为对照组，其余 30 个省份设置为虚拟变量，因此$n=1,2,\cdots,30$，若第i个样本的借款人来自第k个省份，则$Province_k=1$，其余的$Province_i(i\neq k)$全部取 0；若第i个样本来自内蒙古，则所有的$Province_n$全部取0。

在加入控制变量后，本文估计出所有的回归系数，并通过F统计量检验回归系数$\beta_1=\beta_2=\cdots=\beta_{30}$是否为联合显著的，若联合显著，则说明存在地域歧视现象。

为了避免完全多重共线性的影响，引入虚拟变量的个数一般是分类数减1。

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

含有交互项的自变量

price:房价　sqrft:住房面积　bdrms:卧室数量　bthrms:卫生间数量

解释变量对一个因变量

因变量对一个解释变量的偏效应、弹性或半弹性，有时很自然地取决于另一个解释变量的大小。比如，在下式中

$$price=\beta_0+\beta_1 sqrft+\beta_2 bdrms+\beta_3 sqrft\cdot bdrms+\beta_4 bthrms+u$$

bdrms 对 *price* 的偏效应（保持所有其他变量不变）为

$$\frac{\Delta price}{\Delta bdrms}=\beta_2+\beta_3 sqrft \qquad (6.17)$$

若 $\beta_3>0$，则式（6.17）意味着，住房面积越大，增加一间卧室导致价格上升得越多。换言之，住房的平方英尺数与卧室的间数之间存在着**交互效应**（interaction effect）。在总结 *bdrms* 对 *price* 的影响时，我们必须在某些有意义的 *sqrft* 数值（比如样本的均值或上下四分位值）处计算式（6.17）。至于 β_3 是否为零，我们不难检验。

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

回归实例

现有某电商平台846条关于婴幼儿奶粉的销售信息，每条信息由11个指标组成。其中，评价量可以从一个侧面反映顾客对产品的关注度。

请对所给数据进行以下方面的分析，要求最终的分析将不仅仅有益于商家，更有益于宝妈们为宝贝选择适合自己的奶粉。

1) 以评价量为因变量，分析其它变量和评价量之间的关系；

2) 以评价量为因变量，研究影响评价量的重要因素。

	A	B	C	D	E	F	G	H	I	J	K
1	商品名称	商品毛重.k	奶源产地	国产或进口	适用年龄.岁.	包装单位	配方	分类	段位	团购价.元.	评价量
2	美素	1.11	荷兰	进口	1-3岁	桶装	常规配方奶粉	牛奶粉	3段	9.9	683009
3	美素	1.35	荷兰	进口	1-3岁	盒装	常规配方奶粉	牛奶粉	3段	9.9	683009
4	惠氏	1.13	爱尔兰	进口	1-3岁	桶装	常规配方奶粉	牛奶粉	3段	30	605775
5	美素	1.12	荷兰	进口	0.5-1岁	桶装	常规配方奶粉	牛奶粉	2段	28	605775
6	诺优能	0.88	荷兰	进口	3-6岁	桶装	常规配方奶粉	牛奶粉	4段	25.8	605775
7	惠氏	1.16	澳洲/新西兰	国产	1-3岁	桶装	常规配方奶粉	牛奶粉	3段	19.9	605775
8	美赞臣	1.03	荷兰	进口	1-3岁	桶装	常规配方奶粉	牛奶粉	3段	15	605775
9	雅培	1.11	中国大陆	国产	1-3岁	桶装	常规配方奶粉	牛奶粉	3段	36	401183
10	惠氏	1.13	爱尔兰	进口	0.5-1岁	桶装	常规配方奶粉	牛奶粉	1段	36	401183

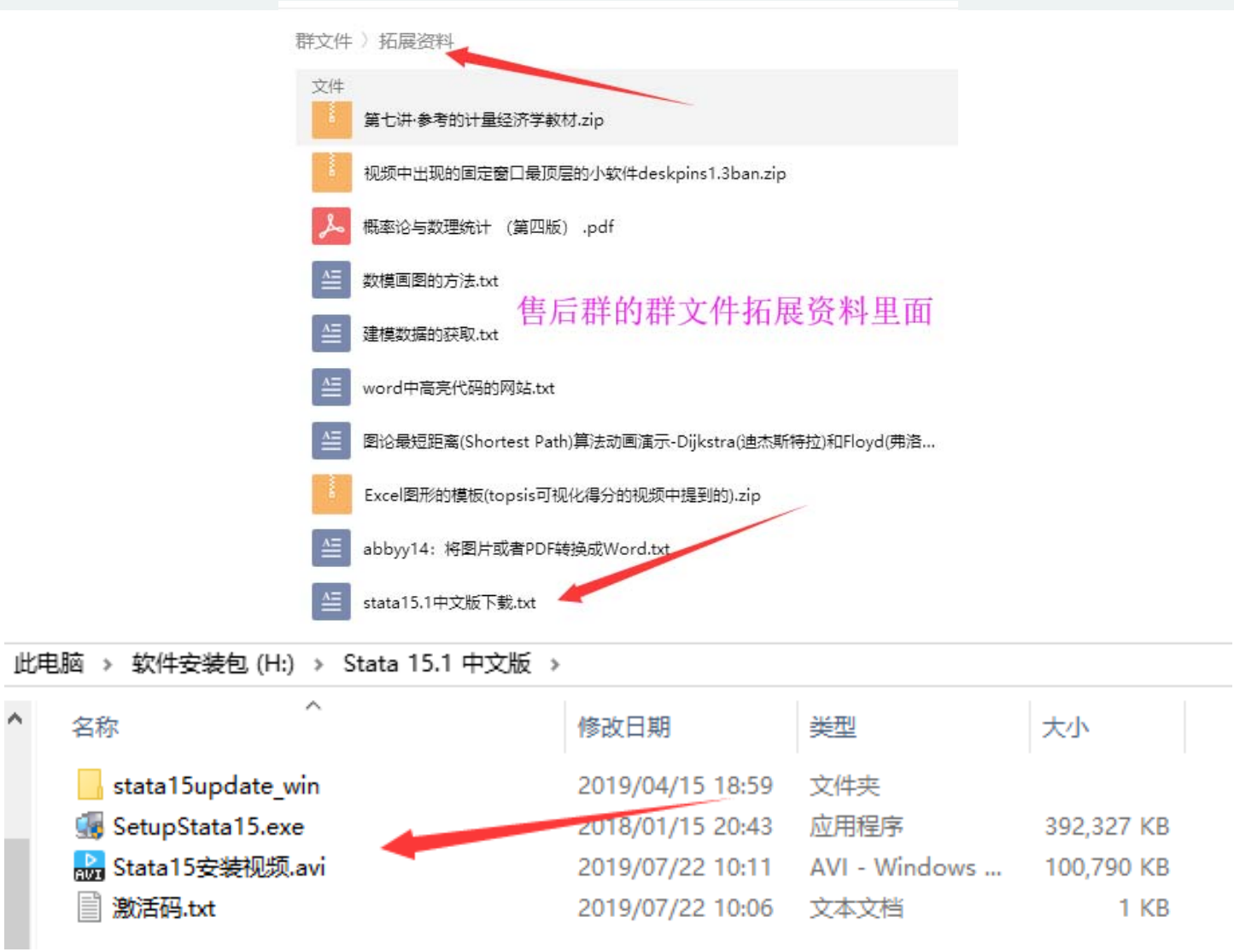
数学建模学习交流

Stata软件介绍

Stata是一个统计分析软件，但它也具有很强的程序语言功能，这给用户提供了一个广阔的开发应用的天地，用户可以充分发挥自己的聪明才智，熟练应用各种技巧，真正做到随心所欲。事实上，Stata的ado文件（高级统计部分）都是用Stata自己的语言编写的。Stata其统计分析能力远远超过了SPSS，在许多方面也超过了SAS！由于Stata在分析时是将数据全部读入内存，在计算全部完成后才和磁盘交换数据，因此计算速度极快（一般来说， SAS的运算速度要比SPSS至少快一个数量级，而Stata的某些模块和执行同样功能的SAS模块比，其速度又比SAS快将近一个数量级！）Stata也是采用命令行方式来操作，但使用上远比SAS简单。其生存数据分析、纵向数据（重复测量数据）分析等模块的功能甚至超过了SAS。用Stata绘制的统计图形相当精美，很有特色。

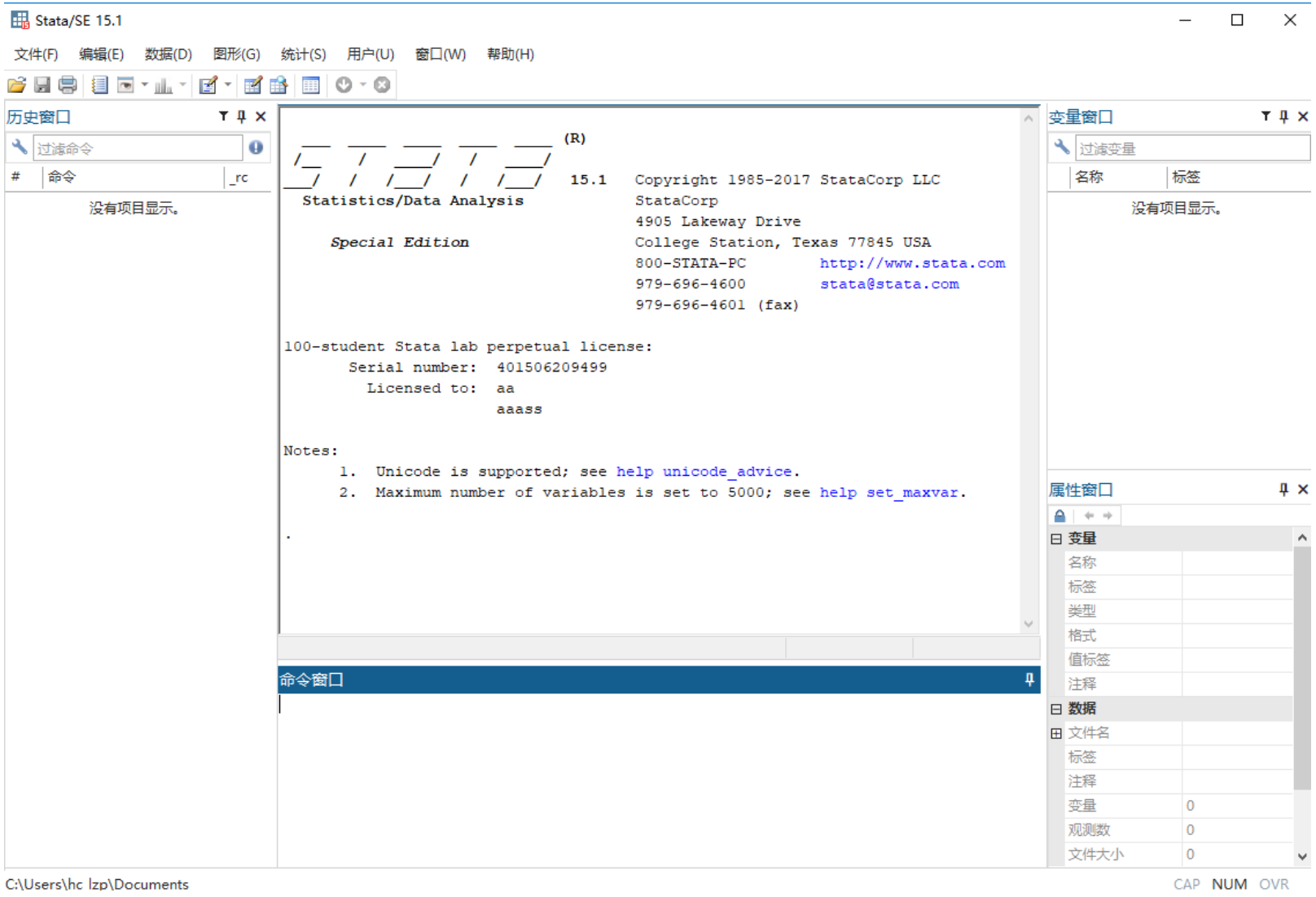
数学建模学习交流

安装方法



数学建模学习交流

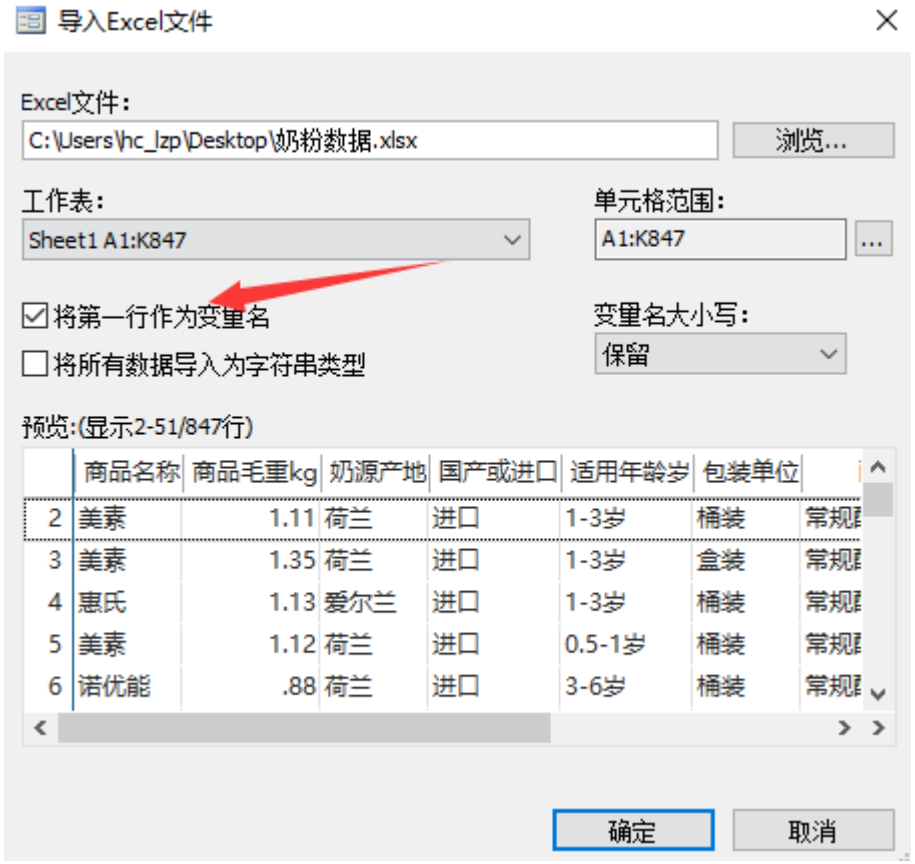
软件界面

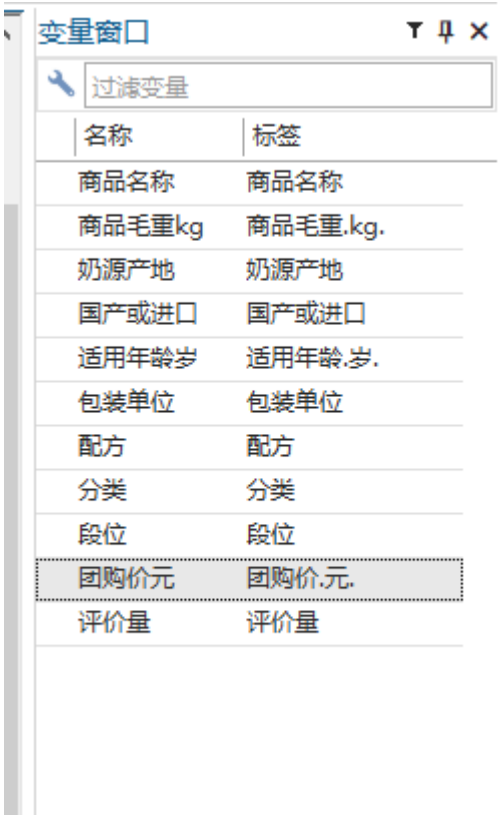


数学建模学习交流

导入数据

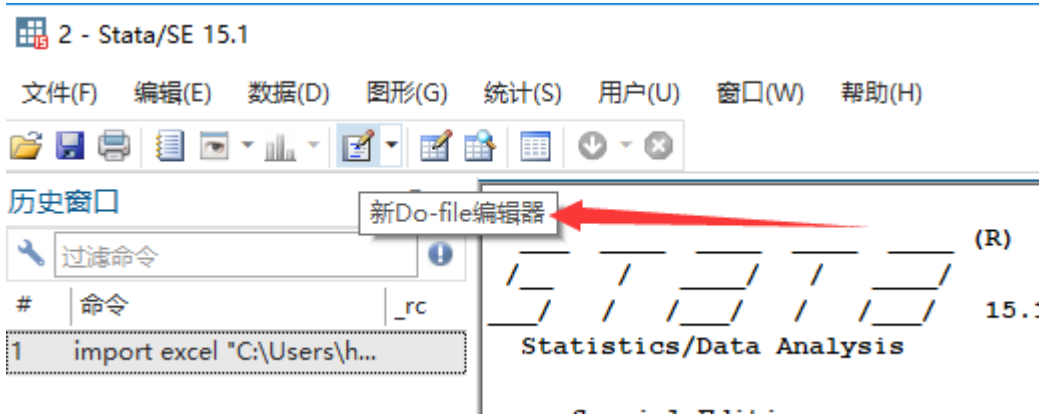
文件 – 导入 – Excel表格

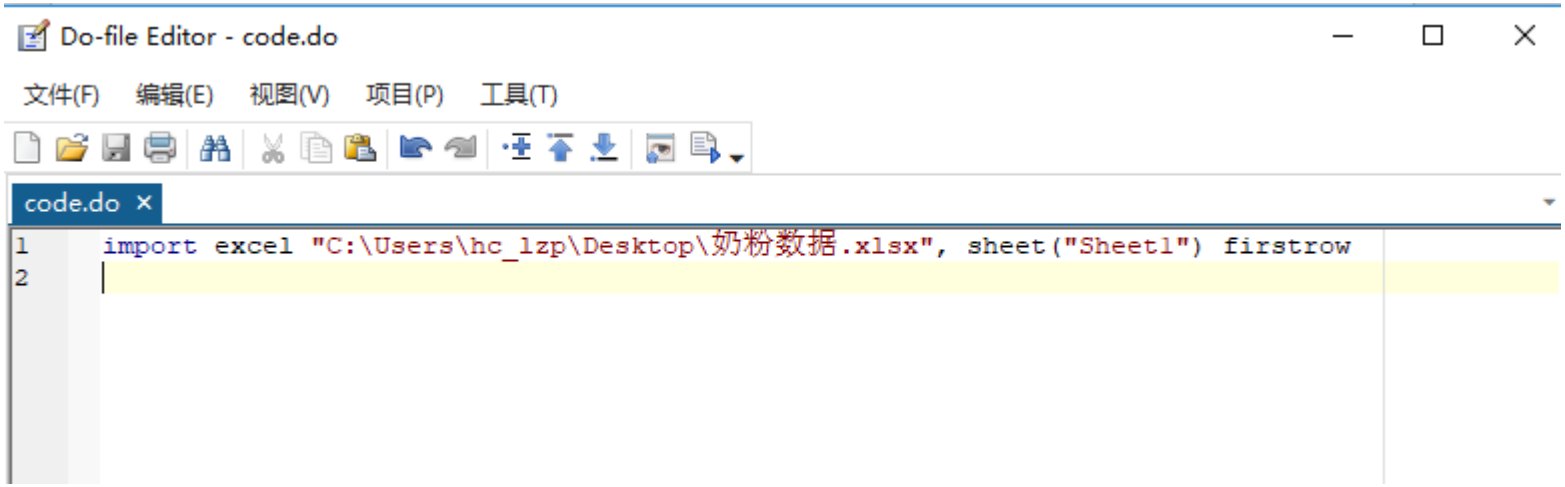




数学建模学习交流

代码记得保存下来





数学建模学习交流

数据的描述性统计

（1）定量数据

summarize 变量1 变量2 ... 变量n

```
. summarize 团购价元 评价量 商品毛重kg
```

Variable	Obs	Mean	Std. Dev.	Min	Max
团购价元	846	366.8944	377.0914	9.9	2598
评价量	846	15800.26	72869.53	1	683009
商品毛重kg	846	1.050684	.7613641	.12	8.64

不要把结果直接截图放到论文中。

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

数据的描述性统计

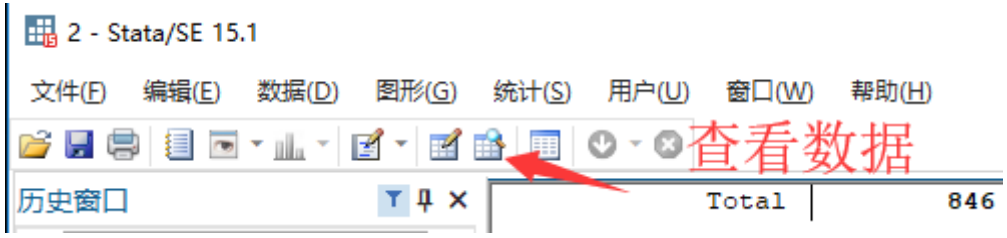
（2）定性数据

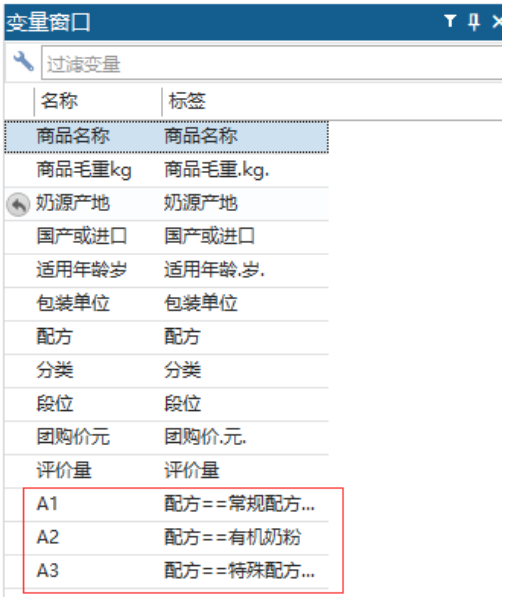
tabulate 变量名,gen(A)

返回对应的这个变量的频率分布表，并生成对应的虚拟变量(以A开头)。

```
. tabulate 配方,gen(A)
```

配方	Freq.	Percent	Cum.
常规配方奶粉	812	95.98	95.98
有机奶粉	24	2.84	98.82
特殊配方奶粉	10	1.18	100.00
Total	846	100.00	

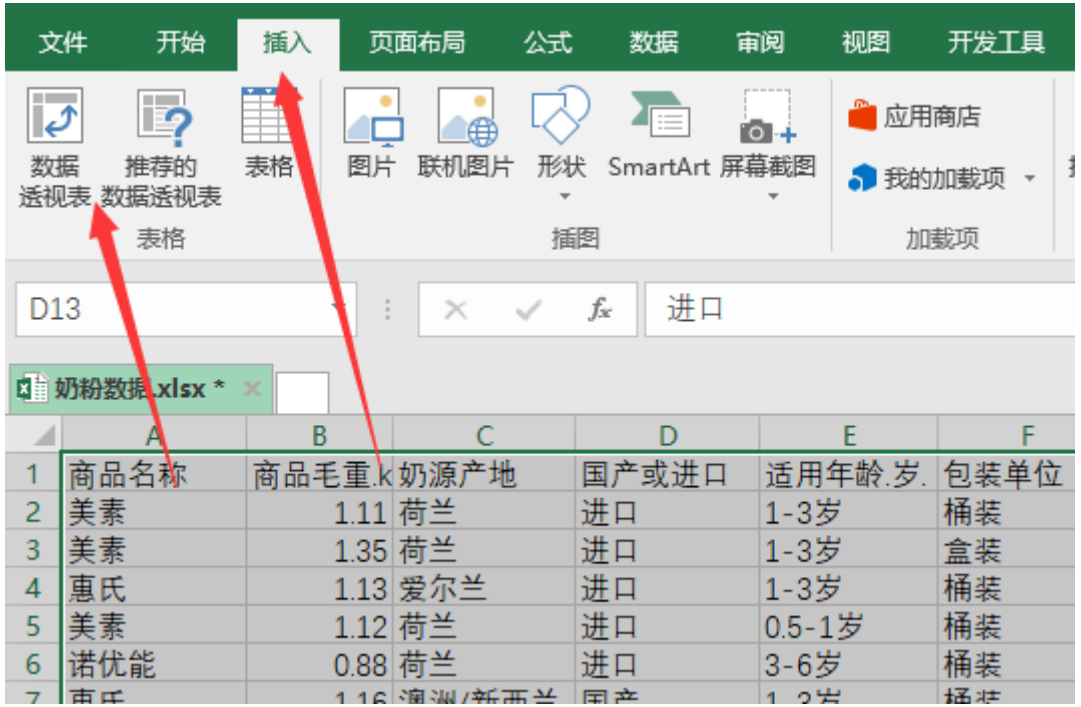




数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

Excel中数据透视表



	A	B	C	D	E	F
1	商品名称	商品毛重.k	奶源产地	国产或进口	适用年龄.岁.	包装单位
2	美素	1.11	荷兰	进口	1-3岁	桶装
3	美素	1.35	荷兰	进口	1-3岁	盒装
4	惠氏	1.13	爱尔兰	进口	1-3岁	桶装
5	美素	1.12	荷兰	进口	0.5-1岁	桶装
6	诺优能	0.88	荷兰	进口	3-6岁	桶装
7	惠氏	1.16	澳洲/新西兰	国产	1-3岁	桶装

视频中演示的都是皮毛
大家可以在课下找找资料系统的学习下。

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

11个指标的总体情况介绍

变量类型	变量名称	说明
定量指标	评价量	间接反映顾客对产品的关注度
	商品毛重（kg）	数据位于0.12-8.64之间
	团购价（元）	数据位于9.9-2598之间
定性指标	商品名称	共有84种不同品牌
	奶源产地	共有9个不同产地
	国产或进口	共有2个类别：进口和出口
	适用年龄.岁.	共有5种类别
	包装单位	共有4种包装单位
	配方	共有三种不同配方
	分类	共有2个类别：牛奶粉和羊奶粉
	段位	共有四种段位，与适用年龄相类似

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

Stata回归的语句

regress y x1 x2 … xk
（默认使用的OLS：普通最小二乘估计法）

```
. regress 评价量 团购价元 商品毛重kg
```

联合显著性检验$\beta_1 = \beta_2 = \cdots = \beta_k = 0$

Source	SS	df	MS		
Model	1.5509e+11	2	7.7543e+10	Number of obs	= 846
Residual	4.3318e+12	843	5.1386e+09	F(2, 843)	= 15.09
				Prob > F	= 0.0000
				R-squared	= 0.0346
				Adj R-squared	= 0.0323
Total	4.4869e+12	845	5.3100e+09	Root MSE	= 71684

评价量	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
团购价元	-35.39873	6.544895	-5.41	0.000	-48.24493	-22.55252
商品毛重kg	2410.303	3241.581	0.74	0.457	-3952.214	8772.821
_cons	26255.38	4909.338	5.35	0.000	16619.42	35891.34

P值小于0.05，代表在95%置信水平下，该回归系数显著的异于0

数学建模学习交流

加入虚拟变量回归

```
. regress 评价量 团购价元 商品毛重kg A1 A2 A3 B1 B2 B3 B4 B5 B6 B7 B8 B9 C1 C2 D1 D2 D3 D4 D5 E1 E
> 2 E3 E4 F1 F2 G1 G2 G3 G4
note: A2 omitted because of collinearity
note: B8 omitted because of collinearity
note: C2 omitted because of collinearity
note: D5 omitted because of collinearity
note: E4 omitted because of collinearity
note: F2 omitted because of collinearity
note: G4 omitted because of collinearity
```

小技巧：在变量窗口中按住Shift不放，可同时选中一列，然后再直接拖动到输入区域

Stata会自动检测数据的完全多重共线性问题。

Source	SS	df	MS		
				Number of obs	= 846
				F(24, 821)	= 3.63
Model	4.3034e+11	24	1.7931e+10	Prob > F	= 0.0000
Residual	4.0566e+12	821	4.9410e+09	R-squared	= 0.0959
				Adj R-squared	= 0.0695
Total	4.4869e+12	845	5.3100e+09	Root MSE	= 70292

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

拟合优度R^2较低怎么办

（1）回归分为解释型回归和预测型回归。
预测型回归一般才会更看重R^2。
解释型回归更多的关注模型整体显著性以及自变量的统计显著性和经济意义显著性即可。
（2）可以对模型进行调整，例如对数据取对数或者平方后再进行回归。
（3）数据中可能有存在异常值或者数据的分布极度不均匀。

补充：关于拟合优度和调整后的拟合优度：

我们引入的自变量越多，拟合优度会变大。但我们倾向于使用调整后的拟合优度，如果新引入的自变量对SSE的减少程度特别少，那么调整后的拟合优度反而会减小。

$$R^2 = 1 - \frac{SSE}{SST} \quad SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

$$R^2_{adjusted} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} \quad (k\text{为自变量的个数})$$

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

标准化回归系数

现有某电商平台846条关于婴幼儿奶粉的销售信息，每条信息由11个指标组成。其中，评价量可以从一个侧面反映顾客对产品的关注度。

请对所给数据进行以下方面的分析，要求最终的分析将不仅仅有益于商家，更有益于宝妈们为宝贝选择适合自己的奶粉。

1) 以评价量为因变量，分析其它变量和评价量之间的关系；

2) 以评价量为因变量，研究影响评价量的重要因素。

为了更为精准的研究影响评价量的重要因素（去除量纲的影响），我们可考虑使用**标准化回归系数**。

对数据进行标准化，就是将原始数据减去它的均数后，再除以该变量的标准差，计算得到新的变量值，新变量构成的回归方程称为标准化回归方程，回归后相应可得到标准化回归系数。

标准化系数的绝对值越大，说明对因变量的影响就越大（只关注显著的回归系数哦）。

数学建模学习交流

Stata标准化回归命令

```
regress y x1 x2 … xk,beta
```

```
. regress 评价量 团购价元 商品毛重kg, b
```

Source	SS	df	MS		
Model	1.5509e+11	2	7.7543e+10	Number of obs	= 846
Residual	4.3318e+12	843	5.1386e+09	F(2, 843)	= 15.09
				Prob > F	= 0.0000
				R-squared	= 0.0346
				Adj R-squared	= 0.0323
Total	4.4869e+12	845	5.3100e+09	Root MSE	= 71684

评价量	Coef.	Std. Err.	t	P>|t|	Beta
团购价元	-35.39873	6.544895	-5.41	0.000	-.1831843
商品毛重kg	2410.303	3241.581	0.74	0.457	.0251836
_cons	26255.38	4909.338	5.35	0.000	.

（1）为什么常数项没有标准化回归系数？

常数的均值是其本身，经过标准化后变成了0.

（2）为啥和之前的回归结果完全相同，除了多了最后那一列标准化回归系数？

对数据进行标准化处理不会影响回归系数的标准误，也不会影响显著性.

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

再来看一道例题

基于多元回归模型的大学生期末数学成绩影响因素分析.pdf

笔者对湖北师范大学某班同学进行了一个学期的观测，收集到该班同学的相关信息，在收集高考成绩时，剔除掉那些不是参加全国卷高考的同学信息；学生大学期间的平时成绩由学生的出勤率、作业情况以及期中检测综合而成，并将得到的数据进行归一化处理，相关数据见下表，其中，是否班干一列中，0 表示非班干，1 表示班干。

表 1　某班成绩归一化数据表

序号	高考数学	高考总分	班干与否	平时成绩	期末成绩
1	0.8125	0.9917	0.0000	0.8947	0.7826
2	0.9375	0.9463	0.0000	1.000	0.9457
3	0.8571	0.9421	1.0000	0.8947	0.7717
4	0.7232	0.9421	0.0000	0.7368	0.5978
5	0.8571	0.9917	1.000	0.8947	0.7065

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

讲讲我的毕业论文

变量	变量说明
SUCCESS	借款是否成功，成功记为1
DEFAULT	获得借款后是否违约，违约记为1
LNAMOUNT	取对数后的借款金额
INTEREST	借款利率
MONTHS	借款的期限，共有6个选择： 3, 6, 9, 12, 18, 24月
INCOME	1表示月收入超过1万元，0表示不超过1万元
HOUSE	有房产则记为1，否则记为0
CAR	有车产则记为1，否则记为0
CREDIT	借款人的信用评级，1表示评级高，0表示评级低
WORKTIME	参加工作的时长，1表示工作时长在3年及以上
MARRED	婚姻状况：已婚记为1，未婚记为0
AGE	借款人的年龄
EDUCATION	本科及以上学历的借款人记为1，低于本科学历记为0

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

扰动项要满足的条件

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \mu_i$$

μ_i为无法观测的且满足一定条件的扰动项

在之前的回归分析中，我们都默认了扰动项是球型扰动项。

球型扰动项：满足“同方差”和“无自相关”两个条件。

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}, \quad Var(\boldsymbol{\mu}|\boldsymbol{X}) = \sigma^2 I_n = \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix}_{n\times n} = \Sigma$$

$\Sigma_{ij} = cov(\mu_i, \mu_j)$，特别地：当$i=j$时，$\Sigma_{ij} = Var(\mu_i) = \sigma^2$

横截面数据容易出现异方差的问题；

时间序列数据容易出现自相关的问题。

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

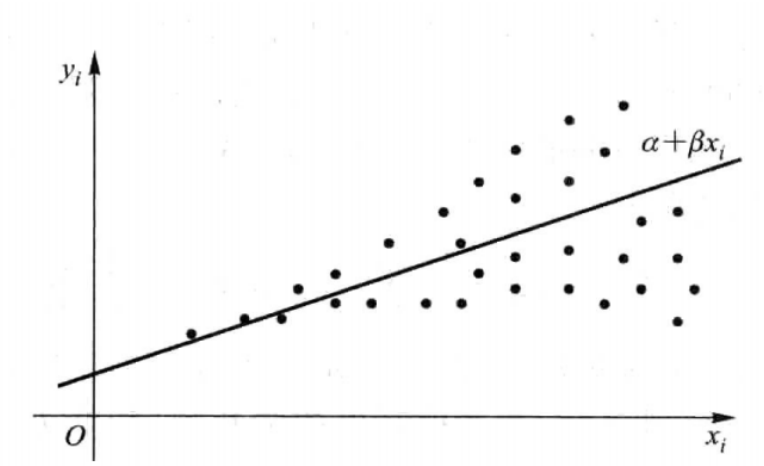
异方差

如果扰动项存在异方差：

（1）OLS估计出来的回归系数是无偏、一致的。

（2）假设检验无法使用（构造的统计量失效了）。

（3）OLS估计量不再是最优线性无偏估计量（BLUE）。



图：异方差的一种类型

注意：这里的信息和熵权法里面确定权重时的信息不是一个意思。异方差这里出现的信息可以理解为对于模型的稳定程度所做的贡献，异方差是指各个扰动项的方差不相同，那么方差较大的扰动项破坏模型稳定性的程度就较大，我们就说它包含的信息量减少。而在熵权法中，方差越大，说明这个指标对于不同个体而言的变化程度就大，那么我们在评价时就不能轻易忽视这个变量。

怎么解决异方差：

（1）使用OLS + 稳健的标准误

（2）广义最小二乘估计法GLS

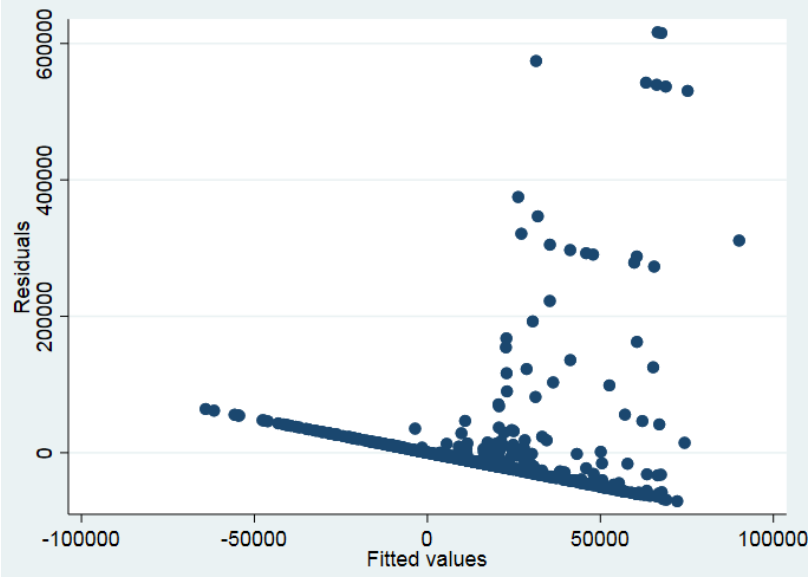
原理：方差较小的数据包含的信息较多，我们可以给予信息量大的数据更大的权重（即方差较小的数据给予更大的权重）

数学建模学习交流

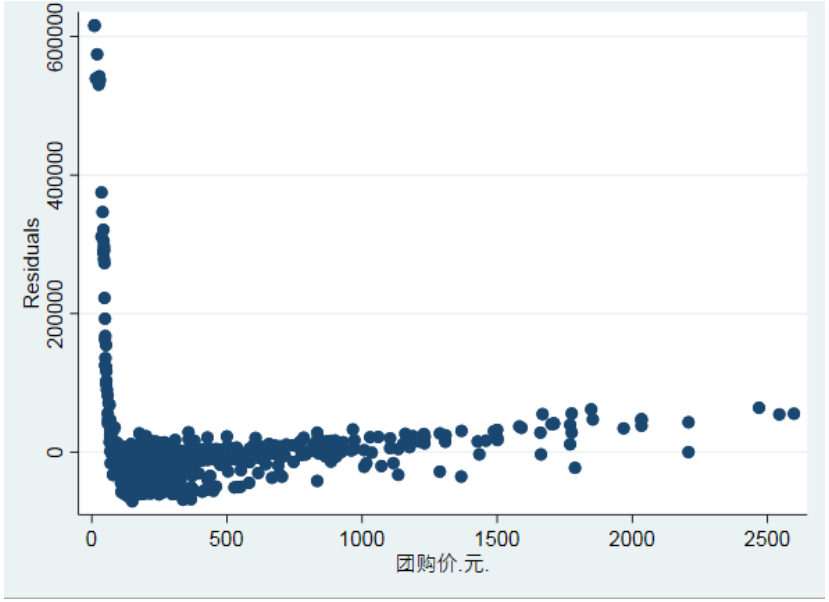
检验异方差

在回归结束后运行命令：

```
rvfplot        (画残差与拟合值的散点图)
rvpplot   x    (画残差与自变量x的散点图)
```



残差与拟合值的散点图



残差与自变量团购价的散点图

注：关于Stata的命令，参考了陈强老师的高级计量经济学

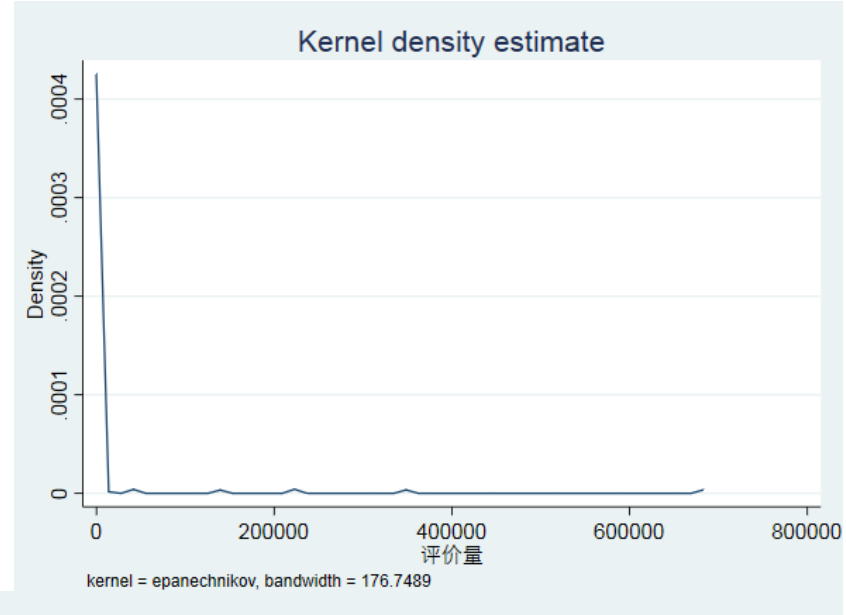
数学建模学习交流

拟合值出现负数的原因

```
. summarize 评价量,d
```

左边百分位数对应的值　　评价量　　最小的四个值

	Percentiles	Smallest		
1%	8	1		
5%	17	2		
10%	31	4	Obs	846
25%	89	4	Sum of Wgt.	846
		最大的四个值		
50%	330.5		Mean	15800.26
		Largest	Std. Dev.	72869.53
75%	1109	605775		
90%	16995	605775	Variance	5.31e+09
95%	52785	683009	Skewness	6.494214
99%	401183	683009	Kurtosis	48.96923



有75%的奶粉品牌的评价量小于1109，评价量超过17000的只有10%不到，而样本均值却达到了15800。这说明评价量的分布极度不平衡，大多数个体的评价量都较小。从右图中也直观的说明了绝大部分品牌的评价量都较小这一特征。

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

异方差的假设检验

3. BP 检验(Breusch and Pagan,1979)

假设回归模型为 $y_i=\beta_1+\beta_2x_{i2}+\cdots+\beta_Kx_{iK}+\varepsilon_i$，检验以下原假设：

$$H_0:\mathrm{E}(\varepsilon_i^2\mid x_2,\cdots,x_K)=\sigma^2 \tag{7.5}$$

如果 H_0 不成立，则条件方差 $\mathrm{E}(\varepsilon_i^2\mid x_2,\cdots,x_K)$ 是 $(x_2,\cdots,x_K)$ 的函数，称为“条件方差函数”(conditional variance function)。BP 检验假设此条件方差函数为线性函数③：

$$\varepsilon_i^2=\delta_1+\delta_2x_{i2}+\cdots+\delta_Kx_{iK}+u_i \tag{7.6}$$

如果认为异方差只与部分解释变量有关，则可以仅使用部分解释变量。也可以添加其他变量，如拟合值 $\hat{y}$，或不在回归方程中的变量 z。根据方程(7.6)，原假设简化为

$$H_0:\delta_2=\cdots=\delta_K=0 \tag{7.7}$$

由于扰动项 ε_i 不可观测，故使用残差平方 e_i^2 对解释变量进行辅助回归：

$$e_i^2=\delta_1+\delta_2x_{i2}+\cdots+\delta_Kx_{iK}+error_i \tag{7.8}$$

仍然使用 nR^2 统计量：

$$nR^2\xrightarrow{d}\chi^2(K-1) \tag{7.9}$$

其中，R^2 为辅助回归的 R^2。BP 检验与怀特检验的区别在于，后者还包括平方项与交叉项。因此，BP 检验可以看成是怀特检验的特例。BP 检验的优点在于其建设性，即可以帮助确认异方差的具体形式。

Stata命令（在回归结束后使用）：

```
estat hettest ,rhs iid
```

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

BP检验的结果

```
.
. // 异方差BP检验
. estat hettest ,rhs iid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: 团购价元 商品毛重kg A1 o.A2 A3 B1 B2 B3 B4 B5 B6 B7 o.B8 B9 C1 o.C2 D1 D2 D3 D4 o.D5 E1 E2 E3 o.E4 F1 o.F2 G1 G2 G3 o.G4

         chi2(24)      =    59.54
         Prob > chi2   =   0.0001

.
```

原假设：扰动项不存在异方差
P值小于0.05，说明在95%的置信水平下拒绝原假设，即我们认为扰动项存在异方差。

数学建模学习交流

怀特检验

2. 怀特检验(White test)

既然在条件同方差下，稳健标准误还原为普通标准误，那么这二者之间的差别就可以用来度量条件异方差。非正式的方法就是用眼睛看一下稳健标准误与普通标准误是否相差不多。怀特(White)1980 年提出的怀特检验正是基于这一思想。

在同方差的原假设 H_0: $E(\varepsilon_i^2 | \boldsymbol{X}) = \sigma^2$ 下，稳健协方差矩阵与普通协方差矩阵之差收敛到一个零矩阵[①]：

$$\hat{\boldsymbol{S}} - s^2 \boldsymbol{S}_{XX} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i' - s^2 \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' = \frac{1}{n}\sum_{i=1}^{n} (e_i^2 - s^2)\boldsymbol{x}_i \boldsymbol{x}_i' \xrightarrow{p} \boldsymbol{0}_{K\times K} \quad (7.1)$$

怀特检验的优点是，它可以检验任何形式的异方差；其缺点则是，如果 H_0 被拒绝，怀特检验并不提供有关异方差具体形式的信息。

```
. estat imtest,white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(151)     =    198.95
         Prob > chi2   =    0.0054

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	198.95	151	0.0054
Skewness	60.66	24	0.0001
Kurtosis	-1.53e+10	1	1.0000
Total	-1.53e+10	176	1.0000

怀特检验原假设：
不存在异方差

Stata命令（在回归结束后使用）：
`estat imtest,white`

数学建模学习交流

异方差的处理方法

如果扰动项存在异方差：
（1）OLS估计出来的回归系数是无偏、一致的。
（2）假设检验无法使用（构造的统计量失效了）。
（3）OLS估计量不再是最优线性无偏估计量（BLUE）。

怎么解决异方差：

（1）使用OLS + 稳健的标准误

如果发现存在异方差，一种处理方法是，仍然进行OLS 回归，但使用稳健标准误。这是最简单，也是目前通用的方法。只要样本容量较大，即使在异方差的情况下，若使用稳健标准误，则所 有参数估计、假设检验均可照常进行。换言之，只要使用了稳健标准误，就可以与异方差“和平共处”了。

（2）广义最小二乘法GLS

原理：方差较大的数据包含的信息较少，我们可以给予信息量大的数据（即方差较小的数据更大的权重）
缺点：我们不知道扰动项真实的协方差矩阵，因此我们只能用样本数据来估计，这样得到的结果不稳健，存在偶然性。

Stock and Watson (2011)推荐，在大多数情况下应该使用“OLS + 稳健标准误”。

数学建模学习交流

使用OLS + 稳健的标准误

```
regress y x1 x2 … xk,robust
```

Linear regression

Number of obs	= 846
F(24, 821)	= 4.49
Prob > F	= 0.0000
R-squared	= 0.0959
Root MSE	= 70292

评价量	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
团购价元	-29.77274	5.522974	-5.39	0.000	-40.61355	-18.93193
商品毛重kg	1557.143	3000.985	0.52	0.604	-4333.363	7447.649
A1	22327.35	10174.21	2.19	0.028	2356.814	42297.88
A2	0	(omitted)				
A3	-5035.106	14050.06	-0.36	0.720	-32613.37	22543.16
B1	24007.69	8596.262	2.79	0.005	7134.454	40880.93
B2	9941.81	6676.533	1.49	0.137	-3163.275	23046.89
B3	38165.85	20521.16	1.86	0.063	-2114.268	78445.96
B4	5554.569	6247.2	0.89	0.374	-6707.795	17816.93
B5	28621.49	9419.918	3.04	0.002	10131.53	47111.45
B6	55533.61	24524.22	2.26	0.024	7396.068	103671.2

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

多重共线性

完全

如果数据矩阵 $\boldsymbol{X}$ 不满列秩(列秩小于 K)，即某一解释变量可以由其他解释变量线性表出，则存在“严格多重共线性”。此时，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$不存在，总体参数 $\boldsymbol{\beta}$ 不可识别，无法定义最小二乘估计量。严格多重共线性在现实数据中很少出现，即使出现，Stata 也会自动识别并删去多余的解释变量。

较常见的是近似(非严格)的多重共线性。其表现为，如果将第 k 个解释变量 x_k 对其余的解释变量 $\{x_1,\cdots,x_{k-1},x_{k+1},\cdots,x_K\}$ 进行回归，所得到的可决系数(记为 R_k^2)较高①。在存在近似多重共线性的情况下，OLS 仍然是最佳线性无偏估计，即在所有线性无偏估计中仍具有最小的方差。但这并不意味着 OLS 估计量方差在绝对意义上小。由于存在多重共线性，矩阵$(\boldsymbol{X}'\boldsymbol{X})$变得几乎不可逆，故从某种意义上来说，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$变得很“大”，致使方差 $\mathrm{Var}(\boldsymbol{b}\mid\boldsymbol{X})=\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$增大，使得对系数的估计变得不准确。在这种情况下，只要数据矩阵 $\boldsymbol{X}$ 中的元素轻微变化，就可能引起$(\boldsymbol{X}'\boldsymbol{X})^{-1}$很大的变化，进而导致 OLS 估计值 $\boldsymbol{b}$ 发生很大变化。通常的“症状”是，虽然整个回归方程的 R^2 较大、F 检验也很显著，但单个系数的 t 检验却不显著，或者系数估计值不合理，甚至符号与理论预期相反。另一可能“症状”是，增减解释变量使得系数估计值发生较大变化(比如，最后加入的解释变量与已有解释变量构成多重共线性)。直观来看，如果两个(或多个)解释变量之间高度相关，则不容易区分它们各自对被解释变量的单独影响力。在极端情况下，一个变量刚好是另一变量的倍数，则完全无法区分。

数学建模学习交流

检验多重共线性

方差膨胀因子(*Variance Inflation Factor*)*VIF*

假设现在有k个自变量，那么第m个自变量的$VIF_m = \frac{1}{1-R^2_{1\sim k\backslash m}}$

$R^2_{1\sim k\backslash m}$是将第$m$个自变量作为因变量，对剩下的$k-1$个自变量回归得到的拟合优度。

VIF_m越大（此时$R^2_{1\sim k\backslash m}$越大），说明第$m$个变量和其他变量的相关性越大。

定义回归模型的$VIF = \max\{VIF_1, VIF_2, \cdots, VIF_k\}$

一个经验规则是，如果$VIF > 10$，则认为该回归方程存在严重的多重共线性。

Variable	VIF	1/VIF
D3	217.74	0.004593
D2	182.09	0.005492
D1	152.17	0.006572
G3	111.03	0.009007
G2	101.57	0.009845
G1	81.69	0.012241
B2	66.48	0.015042
B1	58.34	0.017140
D4	35.10	0.028490
B5	34.05	0.029373
B9	28.93	0.034564
B4	13.19	0.075836
B6	11.29	0.088599
B3	10.25	0.097544
E1	7.71	0.129683
E2	7.15	0.139880
E3	2.09	0.477630
C1	1.86	0.538001
A1	1.78	0.561441
B7	1.66	0.600839
A3	1.54	0.648747
商品毛重kg	1.35	0.743214
F1	1.17	0.851283
团购价元	1.07	0.930527
Mean VIF	47.14	

Stata计算各自变量VIF的命令（在回归结束后使用）：

```
estat  vif
```

数学建模学习交流

多重共线性处理方法

如果发现存在多重共线性，可以采取以下处理方法。

(1)如果不关心具体的回归系数，而只关心整个方程预测被解释变量的能力，则通常可以 不必理会多重共线性（假设你的整个方程是显著的）。这是因为，多重共线性的主要后果是使得对单个变量的贡献估计不准，但所有变量的整体效应仍可以较准确地估计。

(2)如果关心具体的回归系数，但多重共线性并不影响所关心变量的显著性，那么也可以不必理会。即使在有方差膨胀的情况下，这些系数依然显著;如果没有多重共线性，则只会更加显著。

(3) 如果多重共线性影响到所关心变量的显著性，则需要增大样本容量，剔除导致严重共线性的变量（不要轻易删除哦，因为可能会有内生性的影响），或对模型设定进行修改。

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

逐步回归分析

向前逐步回归Forward selection：将自变量逐个引入模型，每引入一个自变量后都要进行检验，显著时才加入回归模型。
(缺点：随着以后其他自变量的引入，原来显著的自变量也可能又变为不显著了，但是，并没有将其及时从回归方程中剔除掉。)

向后逐步回归Backward elimination：与向前逐步回归相反，先将所有变量均放入模型，之后尝试将其中一个自变量从模型中剔除，看整个模型解释因变量的变异是否有显著变化，之后将最没有解释力的那个自变量剔除；此过程不断迭代，直到没有自变量符合剔除的条件。（缺点：一开始把全部变量都引入回归方程，这样计算量比较大。若对一些不重要的变量，一开始就不引入，这样就可以减少一些计算。当然这个缺点随着现在计算机的能力的提升，已经变得不算问题了）

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

Stata实现逐步回归法

向前逐步回归Forward selection：

```
stepwise regress y x1 x2 … xk, pe(#1)
```

pe(#1) specifies the significance level for addition to the model; terms with p<#1 are eligible for addition（显著才加入模型中）.

向后逐步回归Backward elimination：

```
stepwise regress y x1 x2 … xk, pr(#2)
```

pr(#2) specifies the significance level for removal from the model; terms with p>= #2 are eligible for removal（不显著就剔除出模型）.

如果你觉得筛选后的变量仍很多，你可以减小#1或者#2
如果你觉得筛选后的变量太少了，你可以增加#1或者#2

注:
（1）x1 x2 … xk之间不能有完全多重共线性(和regress不同哦)
（2）可以在后面再加参数b和r，即标准化回归系数或稳健标准误

数学建模学习交流

完全多重共线性的错误

```
. stepwise reg 评价量 团购价元 商品毛重kg  A1 A2 A3 B1 B2 B3 B4 B5 B6 B7 B8 B9 C1 C2 D1 D2 D3 D4
> D5 E1 E2 E3 E4 F1 F2 G1 G2 G3 G4,  r pe(0.05)
between-term collinearity, variable A3
r(498);
```

错误原因：出现了完全多重共线性

```
. regress 评价量 团购价元 商品毛重kg A1 A2 A3 B1 B2 B3 B4 B5 B6 B7 B8 B9 C1 C2 D1 D2 D3 D4 D5 E1 E
> 2 E3 E4 F1 F2 G1 G2 G3 G4
note: A2 omitted because of collinearity
note: B8 omitted because of collinearity
note: C2 omitted because of collinearity
note: D5 omitted because of collinearity
note: E4 omitted because of collinearity
note: F2 omitted because of collinearity
note: G4 omitted because of collinearity
```

之前回归时，Stata告诉了我们哪些自变量是完全多重共线性的
（实际上在每个分类变量中任意去除一个元素即可）

```
stepwise reg 评价量 团购价元 商品毛重kg A1 A3 B1 B2 B3 B4 B5
B6 B7 B9 C1 D1 D2 D3 D4 E1 E2 E3 F1 G1 G2 G3, r pe(0.05)
```

数学建模学习交流

向前逐步回归

```
. stepwise reg 评价量 团购价元 商品毛重kg A1 A3 B1 B2 B3 B4 B5 B6 B7 B9 C1 D1 D2 D3 D4 E1 E2 E3 F1 G1 G2 G3,  r pe(0.05)
                      begin with empty model
p = 0.0000 <  0.0500  adding   F1
p = 0.0000 <  0.0500  adding   团购价元
p = 0.0000 <  0.0500  adding   B2
p = 0.0000 <  0.0500  adding   B4
p = 0.0005 <  0.0500  adding   A1
p = 0.0024 <  0.0500  adding   B1
p = 0.0089 <  0.0500  adding   E3
p = 0.0423 <  0.0500  adding   C1

Linear regression                               Number of obs     =        846
                                                F(8, 837)         =       4.37
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0800
                                                Root MSE          =      70227
```

评价量	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
F1	12749.08	3309.649	3.85	0.000	6252.896	19245.27
团购价元	-29.59895	5.305378	-5.58	0.000	-40.01236	-19.18555
B2	-30163.32	6534.212	-4.62	0.000	-42988.69	-17337.96
B4	-37202.15	7989.274	-4.66	0.000	-52883.51	-21520.78
A1	25526.01	7243.579	3.52	0.000	11308.3	39743.72
B1	-15245.51	6567.276	-2.32	0.021	-28135.77	-2355.243
E3	-10468.65	4633.116	-2.26	0.024	-19562.54	-1374.757
C1	-13955.87	6863.629	-2.03	0.042	-27427.82	-483.9274
_cons	16892.02	5423.393	3.11	0.002	6246.978	27537.07

数学建模学习交流

向后逐步回归

```
. stepwise reg 评价量 团购价元 商品毛重kg A1 A3 B1 B2 B3 B4 B5 B6 B7 B9 C1 D1 D2 D3 D4 E1 E2 E3 F1 G1 G2 G3,  r pr(0.05)
                    begin with full model
p = 0.7202 >= 0.0500  removing A3
p = 0.5959 >= 0.0500  removing 商品毛重kg
p = 0.3895 >= 0.0500  removing E3
p = 0.3231 >= 0.0500  removing B4
p = 0.3614 >= 0.0500  removing B7
p = 0.3760 >= 0.0500  removing B2
p = 0.3100 >= 0.0500  removing D4
p = 0.1790 >= 0.0500  removing G2
p = 0.3052 >= 0.0500  removing G3
p = 0.3558 >= 0.0500  removing D3
p = 0.4403 >= 0.0500  removing D2
p = 0.2778 >= 0.0500  removing G1
p = 0.1520 >= 0.0500  removing D1
p = 0.1549 >= 0.0500  removing B3
p = 0.0901 >= 0.0500  removing E1
p = 0.5452 >= 0.0500  removing E2
p = 0.0581 >= 0.0500  removing B6

Linear regression                               Number of obs     =        846
                                                F(7, 838)         =       4.82
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0731
                                                Root MSE          =      70448

------------------------------------------------------------------------------
             |               Robust
      评价量 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      团购价元 |  -30.77121   5.510381    -5.58   0.000    -41.58698   -19.95545
          F1 |   14599.26   4043.991     3.61   0.000     6661.719    22536.8
          A1 |   19848.68   6044.469     3.28   0.001     7984.599    31712.75
          B9 |    38014.6   14988.77     2.54   0.011     8594.661    67434.54
          B1 |   11712.29   4174.431     2.81   0.005     3518.723    19905.86
          B5 |   15240.58   7704.079     1.98   0.048     119.0197    30362.14
          C1 |   -14967.2   5506.036    -2.72   0.007    -25774.44   -4159.956
       _cons |  -4645.301   5377.083    -0.86   0.388    -15199.43    5908.832
------------------------------------------------------------------------------
```

数学建模学习交流

清风"数学建模算法讲解"全套视频，单人购买观看仅需58元

逐步回归的说明

（1）向前逐步回归和向后逐步回归的结果可能不同。

（2）不要轻易使用逐步回归分析，因为剔除了自变量后很有可能会产生新的问题，例如内生性问题。

（3）有没有更加优秀的筛选方法？有的，那就是每种情况都尝试一次，最终一共有$C_k^1 + C_k^2 + \cdots + C_k^k = 2^k - 1$种可能。如果自变量很多，那么计算相当费时。

数学建模学习交流

关注微信公众号"数学建模学习交流"获取更多优质资料

硬核的证明和推导环节

有兴趣的同学请大家看下面这个讲义自学
多元线性回归分析的证明和推导.pdf
(需要很强的线性代数证明基础)

数学建模学习交流

课后作业

（1）如果你看了线性代数的推导pdf，请你用Matlab实现计算回归系数的函数。该函数需要包含以下几个内容：第一：让用户决定是否包含截据项；第二：需要对自变量进行完全多重共线性诊断；第三：如果有能力，你设计的函数可以算出每个自变量对应的标准误，并计算出p值。（可将结果和Stata对照验证你的计算是否正确）

（2）完成论文作业。

（3）对Stata感兴趣的同学可自学人大陈传波Stata十八讲

（4）对经济学建模感兴趣的同学可以自学《计量经济学》

下表是 1990-2007 年中国棉花单产与要素投入表格。请对 5 个要素投入做共线性诊断，并做单产对于 5 个要素投入的逐步回归模型，指出哪个要素投入是最重要的要素？

表 1990-2007 年中国棉花单产与要素投入

年份	单产 kg/公顷	种子费 元/公顷	化肥费 元/公顷	农药费 元/公顷	机械费 元/公顷	灌溉费 元/公顷
1990	1017.0	106.05	495.15	305.1	45.9	56.1
1991	1036.5	113.55	561.45	343.8	68.55	93.3
1992	792.0	104.55	584.85	414	73.2	104.55
1993	861.0	132.75	658.35	453.75	82.95	107.55
1994	901.5	174.3	904.05	625.05	114	152.1
1995	922.5	230.4	1248.75	834.45	143.85	176.4
1996	916.5	238.2	1361.55	720.75	165.15	194.25
1997	976.5	260.1	1337.4	727.65	201.9	291.75
1998	1024.5	270.6	1195.8	775.5	220.5	271.35
1999	1003.5	286.2	1171.8	610.95	195	284.55
2000	1069.5	282.9	1151.55	599.85	190.65	277.35
2001	1168.5	317.85	1105.8	553.8	211.05	290.1
2002	1228.5	319.65	1213.05	513.75	231.6	324.15
2003	1023	368.4	1274.1	567.45	239.85	331.8
2004	1144.5	466.2	1527.9	487.35	408	336.15
2005	1122	449.85	1703.25	555.15	402.3	358.8
2006	1276.5	537	1888.5	637.2	480.75	428.4
2007	1233	565.5	2009.85	715.65	562.05	456.9

数学建模学习交流